AD________________

Award Number: DAMD17-00-1-0050

TITLE: The Prostate Expression Database

PRINCIPAL INVESTIGATOR: Peter S. Nelson, M.D.

CONTRACTING ORGANIZATION: Fred Hutchinson Cancer Research Center
Seattle, WA 98104-2092

REPORT DATE: April 2002

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 074-0188*

1. AGENCY USE ONLY (Leave blank)	**2. REPORT DATE** April 2002	**3. REPORT TYPE AND DATES COVERED** Annual (1 Apr 01 - 31 Mar 02)

4. TITLE AND SUBTITLE The Prostate Expression Database	**5. FUNDING NUMBERS** DAMD17-00-1-0050
6. AUTHOR(S) Peter S. Nelson, M.D.	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Fred Hutchinson Cancer Research Center Seattle, WA 98104-2092 **E-Mail:** pnelson@fhcrc.org	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited	**12b. DISTRIBUTION CODE**

13. ABSTRACT *(Maximum 200 Words)*

This proposal aims to exploit advances in biotechnology and informatics to develop a genetics resource termed the Prostate Expression Database (PEDB) (http://www.pedb.org). . The foundation of PEDB is the identification and characterization of a prostate transcriptome, the intermediary between the genome and the proteome that represents that portion of the human genome actively used or transcribed in the prostate. The research accomplished to date has assembled a working virtual prostate transcriptome that defines the genes used or transcribed in prostate cell types and tissues. This transcriptome has a physical counterpart of 6,500 cDNAs arrayed in cDNA microarray format for large-scale expression studies. This transcriptome has been used as a foundation for studies of the prostate proteome, the working counterpart to the genome and transcriptome. Our results show that these approaches are complementary. Analysis of the virtual transcriptome of LNCaP cells has identified 15 new androgen-regulated genes to date. Characterization of these genes is in progress. We have extended PEDB to include sequence analysis of the murine prostate and constructed a corresponding database, mPEDB to facilitate the dissemination of mouse prostate gene expression information.

14. SUBJECT TERMS prostate cancer, transcriptome, cDNA, database	**15. NUMBER OF PAGES** 24
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

___ Where copyrighted material is quoted, permission has been obtained to use such material.

___ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

___ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

NA ~~X~~ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

4-30-02

PI - Signature Date

TABLE OF CONTENTS **PAGE**

INTRODUCTION

This proposal aims to exploit advances in biotechnology and informatics to develop a genetics resource termed the Prostate Expression Database (PEDB) (http://www.pedb.org). PEDB is an integrated resource focused exclusively on prostate cancer that incorporates public DNA and protein sequence and informatics resources where applicable. The foundation of PEDB is the identification and characterization of a prostate transcriptome, the intermediary between the genome and the proteome that represents that portion of the human genome actively used or transcribed in the prostate.

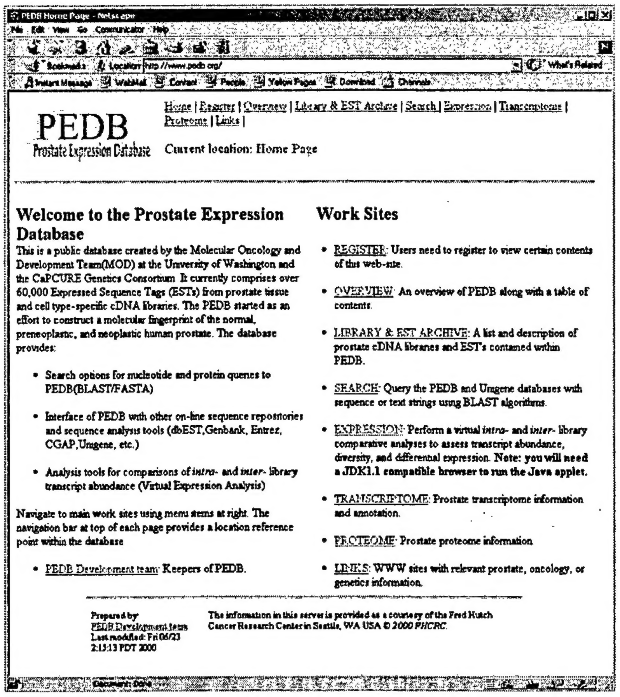
This proposal extends the PEDB capabilities by accomplishing the following specific objectives:
1) assemble and annotate a working prostate transcriptome;
2) develop a suite of database tools to facilitate investigator-initiated database queries;
3) extend the prostate transcriptome in 3 dimensions: acquiring rare transcripts, assembling sequences representing full-length genes, and mapping the locations of interesting and novel prostate genes;
4) assemble a solid-phase nonredundant archive of prostate-derived cDNA clones for distribution to investigators and to the Image Consortium sites.

BODY

The following summarizes the technical objectives for the proposal and the work accomplished during the 12-month interval between the last report (03/14/01) and the preparation of this report (03/15/02).

Technical objective 1: *To assemble and annotate a working prostate transcriptome* (months 1-16)

- *Task 1: Install Phrap, d2-cluster, and CAP3 software and test on small, known genomic sequence assemblies (months 1-3).* Completed (Prior Progress Report). Phrap is the selected assembly algorithm of choice.
- *Task 2: Assemble UniGene and prostate EST test sets. Compare with previous assemblies performed with CAP2. Manually review assembly discrepancies. Compare assemblies with UniGene and CGAP clusters (months 1-6).* Completed (Prior Progress Report)..
- *Task 3: Assemble and annotate all PEDB sequences using best available algorithm (months 6-12).* Completed. We have completed the download of 20,000 additional chromatograms from the Washington University web site and added an additional 5,000 ESTs from our own sequencing project. The assembly of these ESTs with phrap has been completed and the contigs have been annotated against sequences housed in Genbank and Unigene. The latest assembly statistics are shown in Figures 1 and 2 with the assembly schema and results.


PEDB Home Page - Netscape
File Edit View Go Communicator Help
Bookmarks Location http://www.pedb.org/ What's Related
Instant Message WebMail Contact People Yellow Pages Download Channels
Document: Done

PEDB
Prostate Expression Database

Home | Register | Overview | Library & EST Archive | Search | Expression | Transcriptome | Proteome | Links |

Current location: Home Page

Welcome to the Prostate Expression Database

This is a public database created by the Molecular Oncology and Development Team(MOD) at the University of Washington and the CaPCURE Genetics Consortium. It currently comprises over 60,000 Expressed Sequence Tags (ESTs) from prostate tissue and cell type-specific cDNA libraries. The PEDB started as an effort to construct a molecular fingerprint of the normal, preneoplastic, and neoplastic human prostate. The database provides:

- Search options for nucleotide and protein queries to PEDB(BLAST/FASTA)
- Interface of PEDB with other on-line sequence repositories and sequence analysis tools (dbEST,Genbank, Entrez, CGAP,Unigene, etc.)
- Analysis tools for comparisons of *intra*- and *inter*- library transcript abundance (Virtual Expression Analysis)

Navigate to main work sites using menu items at right. The navigation bar at top of each page provides a location reference point within the database

- PEDB Development team: Keepers of PEDB.

Work Sites

- REGISTER: Users need to register to view certain contents of this web-site.
- OVERVIEW: An overview of PEDB along with a table of contents.
- LIBRARY & EST ARCHIVE: A list and description of prostate cDNA libraries and EST's contained within PEDB.
- SEARCH: Query the PEDB and Unigene databases with sequence or text strings using BLAST algorithms.
- EXPRESSION: Perform a virtual *intra*- and *inter*- library comparative analyses to assess transcript abundance, diversity, and differential expression. **Note: you will need a JDK1.1 compatible browser to run the Java applet.**
- TRANSCRIPTOME: Prostate transcriptome information and annotation.
- PROTEOME: Prostate proteome information
- LINKS: WWW sites with relevant prostate, oncology, or genetics information.

Prepared by
PEDB Development team
Last modified: Fri 06/23
2:15:13 PDT 2000

The information in this server is provided as a courtesy of the Fred Hutch Cancer Research Center in Seattle, WA USA © *2000 FHCRC.*

55 cDNA libraries
•normal
•epithelium
•stroma
•PIN
•primary CA
•metastatic CA
•cell lines
•xenografts
↓
~110,000 ESTs
↓
Assembled
Annotated
↓
Prostate Transcriptome

Figure 1. (above) WWW Interface for the Prostate Expression Database (PEDB). 55 libraries containing **~110,000 ESTs** have been processed, assembled, and annotated to comprise a Prostate Transcriptome.

- *Task 4: Develop a gene classification schema based upon function, and automate assignments of clusters to functional groups (months 8-16).* We have completed a functional annotation scheme modeled after the TIGR annotation scheme for partitioning genes into cellular functional groups. This has been applied to the LNCaP dataset and is available for viewing and analysis on the PEDB website(Prior Progress Report).
- *Task 5: Develop a Graphical User Interface for viewing and navigating between sequence, functional group, and expression data (months 3-12).* A user interface has been developed for the viewing of sequence chromatograms and for searching the database with keywords in addition to BLAST queries (Prior Progress Report)..
- *Task 6: Write scripts to automate the input of new prostate ESTs, processing of new ESTs, clustering of the database sequences and annotation of the entire cluster complement on a monthly basis. (months 12-16*). Completed. (Prior Progress Report).

<u>Technical objective 2:</u> *To develop a suite of database tools to facilitate investigator-initiated database queries. (months 1-18).* In progress. We have completed an improved search engine to incorporate searches of sequences with truncated key words and wildcards.

- *Task 7: Evaluate potential sequence cluster/assembly viewing tools: DrawMap, Consed, Phrapview, CloneView, and AlignView (months 6-12).* We have completed the evaluation of sequence assembly viewing tools and have selected Consed as the viewer of choice.
- *Task 8: Design a client application (ContigView) extending the functionality of the Virtual Expression Analysis Tool to view the cluster output produced by the best algorithm as identified in specific aim 1. (months 8-14).* In progress.
- *Task 9: Design GUI to support high level viewing of clustered data with graphical maps incorporating zoom features for viewing nucleotide sequence traces and assemblies (8-14).* A tool for viewing individual sequence traces has been developed and implemented into PEDB (Prior Progress Report). A tool for viewing sequence assemblies is in progress.
- *Task 10: Write Java code for ContigView and test on datasets representing assemblies of few and many ESTs with both short and long consensus sequences. (months 10-24)* We have opted not to incorporate Java code for this process. Current software is being developed in C and PERL.
- *Task 11: Test (and modify if necessary) ContigView on Windows/NT/MacIntosh/Unix operating systems (months 24-30).* In progress.
- *Task 12: Write applications to link cluster consensus to relevant public databases (Genbank, etc) (months 20-24).* In progress.
- *Task 13: Write applications for integrating gene analysis tools: exon prediction, promoter finders, transcription factor binding site ID, protein motif ID (months 18-28).* We are currently testing different exon prediction programs. We are exploring the utility of incorporating the GUI used by the SantaCruz human genome assembly site for representing PEDB assemblies (http://genome.ucsc.edu/).
- *Task 14: Evaluate the incorporation of software for SNP detection (PolyPhred) in client-selected PEDB clusters (26-30).* We have evaluated the utility of incorporating EST-derived SNPs. After discussions with Dr. Debbie Nickerson (University of Washington), the accuracy of EST-derived SNPs, and hence their utility for polymorphism detection, requires further study. We are currently performing a comparative analysis of ESTs derived through our sequencing efforts, with corresponding ESTs derived from the same gene to assess the ultimate utility of these sequences.

Technical objective 3: To *extend the prostate transcriptome in 3 dimensions: 1) acquire rare transcripts 2) assemble sequences representing full-length genes and 3) map the location to EST clusters to specific chromosomal sites.* (months 12-25)

- *Task 15: construct LNCaP random primed library and CAP-finder library (months 6-7).* We have constructed one prostate cDNA library from androgen stimulated LNCaP and one cDNA library from androgen-starved LNCaP and compared these expression profiles to identify androgen-regulated genes in prostate epithelium(see Clegg et al in reportable outcomes). We have constructed one prostate cDNA library from prostate small cell carcinoma and sequenced approximately 3,500 ESTs. Virtual comparison's of this library against prostate adenocarcinoma and small cell lung carcinoma has identified several known genes and novel sequences that may be useful for studying the development, progression, and therapy of this variant of prostate cancer (see Clegg et al in reportable outcomes)

- *Task 16: partially sequence 1,600 cDNAs from each library and enter ESTs into PEDB. (months 8-12)* See above. Approximately 2,000 additional ESTs from the LNCaP libraries and 3,500 ESTs from the prostate small cell library have been entered into PEDB, assembled using phrap, and annotated against sequences present in the public nucleotide databases.
- *Task 17: as above with normal prostate tissue (months 13-18).* We have constructed cDNA libraries from microdissected luminal cell, basal cell, and stromal tissue. A total of 2,200 ESTs have now been produced from these libraries.
- *Task 18: as above with microdissected primary prostate cancer tissue (months 25-30).* Two libraries representing primary prostate carcinoma have now been constructed. Quality assessment is in progress.
- *Task 19: "Negative Select" 10,000 cDNAs from normal prostate cDNA array (months 19-20).* We have now selected ~6,500 non-redundant clones for array construction.
- *Task 20: partially sequence 10,000 negatively selected, low abundance cDNAs and submit ESTs into PEDB (months 21-25).* The selection of low abundance clones is in progress.
- *Task 21: Identify 60 interesting uncharacterized prostate ESTs/cDNAs based upon a) homology to known physiologically important genes or b) novelty, to directly obtain full-length cDNA sequence using RACE, library screening, genomic assembly, and primer-directed sequencing. A total of 15 full-length cDNAs per year will be obtained (ongoing throughout period of award).* Since the previous progress report, we have cloned and sequenced the full-length of 4 androgen-regulated prostate cDNAs; PWDMP, PART2, 6A4, and KIAA0056. Characterization of these genes is in progress.
- *Task 22: Map interesting prostate cDNAs described above using radiation hybrid panel mapping. (ongoing throughout period of award).* With the completion of the human genome project, this aim is essentially obsolete. We are confirming the genomic location for selected genes.
- *Task 23: submit data to PEDB and public databases* (ongoing throughout period of award).

Technical objective 4: *To assemble a solid phase nonredundant archive of prostate-derived cDNA clones.*

- *Task 24: identify a cohort consisting of 3,000 distinct, unique prostate clusters from year 1 PEDB assembly (month 10).* We have assembled a non-redundant set of 6,500 cDNAs from LNCaP, normal and neoplastic prostate cDNA libraries. These have now been re-arrayed into 96-well and 384-well microtiter plates. The clone set has been replicated. PCR amplification has been performed. In addition to this effort, we have initiated an effort to assemble a working transcriptome of the murine prostate gland. To this end we have constructed 10 mouse prostate cDNA libraries, sequenced >16,000 ESTs, and assembled a non-redundant clone set of >4,000 cDNAs. Sequence verification is currently in progress. This will be followed by the construction of a mouse prostate cDNA array.
- *Task 25: cross-reference cluster sequences with PEDB clone archive to determine the clones physically available for biological studies (month 11).* Completed.
- *Task 26 determine the longest physical clone for each cluster and consolidate bacterial transformants into 96-well plates using the Genetix Q-bot. Preserve for storage (months 12-13).* Completed.

- *Task 27: annotate and ship to IMAGE consortium clone distributors (month 14).* In progress.
- *Task 28: repeat Tasks 24-27 for 3,000 additional unique clusters at the end of month 24.* Completed.
- *Task 29: repeat Tasks 24-27 for 3,000 additional unique clusters at the end of month 35.* In Progress.
- *Task 30: plan for incorporation and integration of PEDB with microarray data and proteomics data (months 24-36).* Currently in planning stages. We have obtained database software for archiving and analyzing microarray data from Stanford University. We are currently testing for compatibility with PEDB.
- *Task 31: analyze/compile data and prepare formal report (month 36).*

KEY RESEARCH ACCOMPLISHMENTS

- Selected phrap as the sequence assembly algorithm for PEDB. Assembled and annotated 110,000 PEDB ESTs.
- Constructed cDNA libraries from microdissected luminal epithelial cells, basal epithelial cells, stromal cells, and primary prostate carcinoma.
- Sequenced cDNAs from LNCaP, luminal cell, basal cell and stromal cell cDNA libraries (total ~10,000 ESTs) and assembled the ESTs into clusters/contigs. The data indicate the libraries are of good quality with significant diversity.
- Virtual comparison of the LNCaP libraries identified 4 additional new androgen-regulated genes. Northern analysis confirmed androgen-regulation for these genes.
- Constructed a cDNA library of prostate small cell carcinoma. 3,000 cDNA clones have been sequenced and deposited into PEDB.
- Compiled a non-redundant virtual and physical archive of prostate ESTs/cDNAs comprising 6,500 distinct species. These clones have been consolidated, replicated, and arrayed for cDNA microarray analysis.
- Initiated the characterization of the mouse prostate transcriptome. Constructed 10 mouse prostate cDNA libraries and produced >16,000 ESTs.

REPORTABLE OUTCOMES

Clegg N, Eroglu B, Ferguson C, Arnold H, Moorman A, **Nelson PS**. Digital expression profiles of the prostate androgen-response program. J Steroid Biochem Mol Biol. 2002 Jan;80(1):13-23.

Liu AY, **Nelson PS**, van den Engh G, Hood L. Human prostate epithelial cell-type cDNA libraries and prostate expression patterns. Prostate. 2002 Feb 1;50(2):92-103.

Nelson PS, Pritchard C, Abbott D, Clegg N. The human (PEDB) and mouse (mPEDB) Prostate Expression Databases. Nucleic Acids Res. 2002 Jan 1;30(1):218-20.

Clegg N, Eroglu B, Ferguson C, Arnold H, Moorman A, True L, Vessella R, and **Nelson PS**, *Transcript analysis of prostate small cell carcinoma.* (Submittted) Prostate.

CONCLUSIONS

The research accomplished to date has assembled a working virtual prostate transcriptome that defines the genes used or transcribed in prostate cell types and tissues. This transcriptome has a physical counterpart of 6,500 cDNAs arrayed in cDNA microarray format for large-scale expression studies. This transcriptome has been used as a foundation for studies of the prostate proteome, the working counterpart to the genome and transcriptome. Our results show that these approaches are complementary. Analysis of the virtual transcriptome of LNCaP cells has identified 15 new androgen-regulated genes to date. Characterization of these genes is in progress. We have extended the human PEDB to also encompass murine prostate gene expression in mPEDB. We anticipate that mPEDB will be a valuable resource for research involving murine models of prostate carcinoma. Immediate uses for mPEDB involve comparative gene expression studies with the human prostate.

REFERENCES

None

APPENDICES

Nelson PS, Pritchard C, Abbott D, Clegg N. The human (PEDB) and mouse (mPEDB) Prostate Expression Databases. Nucleic Acids Res. 2002 Jan 1;30(1):218-20.

Clegg N, Eroglu B, Ferguson C, Arnold H, Moorman A, **Nelson PS**. Digital expression profiles of the prostate androgen-response program. J Steroid Biochem Mol Biol. 2002 Jan;80(1):13-23.

PERGAMON

Journal of Steroid Biochemistry & Molecular Biology 80 (2002) 13–23

The Journal of
Steroid Biochemistry
&
Molecular Biology

www.elsevier.com/locate/jsbmb

Digital expression profiles of the prostate androgen-response program

Nigel Clegg, Burak Eroglu, Camari Ferguson, Hugh Arnold, Alec Moorman, Peter S. Nelson*

Division of Human Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109, USA

Received 15 June 2001; accepted 24 September 2001

Abstract

The androgen receptor (AR) and cognate ligands regulate vital aspects of prostate cellular growth and function including proliferation, differentiation, apoptosis, lipid metabolism, and secretory action. In addition, the AR pathway also influences pathological processes of the prostate such as benign prostatic hypertrophy and prostate carcinogenesis. The pivotal role of androgens and the AR in prostate biology prompted this study with the objective of identifying molecular mediators of androgen action. Our approach was designed to compare transcriptomes of the LNCaP prostate cancer cell line under conditions of androgen depletion and androgen stimulation by generating and comparing collections of expressed sequence tags (ESTs). A total of 4400 ESTs were produced from LNCaP cDNA libraries and these ESTs assembled into 2486 distinct transcripts. Rigorous statistical analysis of the expression profiles indicated that 17 genes exhibited a high probability ($P > 0.9$) of androgen-regulated expression. Northern analysis confirmed that the expression of *KLK3/PSA*, *FKBP5*, *KRT18*, *DKFZP564K247*, *DDX15*, and *HSP90* is regulated by androgen exposure. Of these, only *KLK3/PSA* is known to be androgen-regulated while the other genes represent new members of the androgen-response program in prostate epithelium. LNCaP gene expression profiles defined by two independent experiments using the serial analysis of gene expression (SAGE) method were compared with the EST profiles. Distinctly different expression patterns were produced from each dataset. These results are indicative of the sensitivity of the methods to experimental conditions and demonstrate the power and the statistical limitations of digital expression analyses. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Androgen; Prostate; EST; SAGE; Transcriptome

1. Introduction

Genes regulated by androgenic hormones are of critical importance for the normal physiological function of the human prostate gland, and they contribute to the development of prostate diseases such as benign prostatic hypertrophy (BPH) and prostate carcinoma. Androgens such as testosterone and dihydrotestosterone (DHT) interact with the androgen receptor (AR) leading to the transcriptional activation of androgen-target genes [1]. This gene network regulates prostate morphogenesis, growth, and function, and promotes the development and progression of prostate neoplasia [2]. Despite the importance of androgens in modulating diverse prostate cellular processes, relatively few components of this androgen-response program have been identified or characterized.

Current estimates indicate that between 35,000 and 40,000 genes are encoded in the human genome [3,4]. To confer developmental and functional specificity, only a fraction of this total is transcribed in a given tissue or cell type at any given time. This repertoire of expressed genes in transcript form is termed the transcriptome [5], a dynamic assessment or inventory of gene expression activity that reflects the cellular developmental state and response(s) to environmental perturbations. Proceeding from the hypothesis that comprehensive gene expression profiles will provide insights into cellular function, several procedures have been developed to qualitatively and quantitatively assess transcriptomes. These methods can be broadly divided into analog approaches

Abbreviations: KLK3, kallikrein 3; *RPLP0*, ribosomal protein large, P0; *UQCRC2*, ubiquinol-cytochrome *c* reductase core protein 2; *FKBP5*, FK506-binding protein 5; *DKFZP564K247*, DKFZP564K247 protein; *PHGDH*, phosphoglycerate dehydrogenase; *KRT18*, keratin 18; *RPS25*, ribosomal protein S25; *E1F3S6*, eukaryotic translation initiation factor 3, subunit 6 (48 kDa); *FTL*, ferritin, light polypeptide; *DDX15*, DEAD/H (Asp-Glu-Ala/His) box polypeptide; *RPS27A*, ribosomal protein S27A; *ACADVL*, acyl-coenzyme A dehydrogenase, very long chain; *KIAA0101*, KIA0101 gene product; *DKFZP564D0462*, hypothetical protein DKFZP-564D0462; *RPS15A*, ribosomal protein S15a; *DED*, apoptosis antagonizing transcription factor; *BSG*, basigin; *TPI1*, triosephosphate isomerase 1; *CLTB*, clathrin, light polypeptide (Lcb); *DBI*, diazepam binding inhibitor; *ENO1*, enolase 1 (alpha); *KLK2*, kallikrein 2; *KLK4*, kallikrein 4; *ODC1*, ornithine decarboxylase 1; *PDHA1*, pyruvate dehydrogenase (lipoamide) alpha 1; *TMEPA1*, transmembrane, prostate androgen-induced RNA; *TUBA1*, tubulin, alpha 1; *UGT2B17*, UDP glycosyltransferase 2 family, polypeptide B17; *VEGF*, vascular endothelial growth factor

* Corresponding author. Fax: +1-206-667-2917.
E-mail address: pnelson@fhcrc.org (P.S. Nelson).

0960-0760/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S0960-0760(01)00167-4

such as DNA array analysis [6–8], and digital methods as exemplified by expressed sequence tag (EST) quantitation [9] and the serial analysis of gene expression (SAGE) [10]. Each approach has distinct advantages and limitations that have been detailed previously [11]. A principle advantage of digital methods is the possibility of sampling the complete transcriptome in a single experiment. These approaches also permit the analysis of previously uncharacterized genes and allow for direct statistical analyses of transcript numbers rather than relying on indirect measures of transcript ratios.

Our objective in this study was to identify genes expressed in human prostate cells exhibiting transcriptional regulation by androgens. We hypothesize that such genes could be direct mediators of the androgen-receptor pathway or be involved in prostate-specific functions that could be exploited for understanding normal and neoplastic prostate growth. To facilitate systematic studies of prostate gene expression, we have established the prostate expression database (PEDB), an archive that contains more than 70,000 ESTs generated from prostate cDNA libraries [12]. Two libraries constructed specifically for this study comprise genes expressed in the LNCaP prostate cancer cell line under conditions of androgen stimulation and androgen deprivation. The LNCaP cell line represents a model system for the study of androgen regulation as LNCaP cells express a functional AR, proliferate in response to physiological levels of androgens, and increase the transcription of known androgen-regulated genes such as prostate specific antigen (PSA) [13]. We applied statistical tools to compare these EST datasets and identified both known and novel genes with a high probability ($P > 0.9$) of being regulated by androgens. Northern analysis was used to confirm androgen-regulated expression. These studies identified *FKBP5*, *KRT18*, *DKFZP564K247*, *DDX15*, and *HSP90*, as new members of the prostate epithelial androgen-response program. LNCaP transcriptomes defined by two distinct SAGE experiments were also examined for genes exhibiting androgen regulation and these results were compared with the EST profiles. These results support the use of comprehensive gene expression profiling methods to define cellular responses to hormonal stimuli, and demonstrate both the power and the statistical limitations of digital expression analyses.

2. Materials and methods

2.1. Cell culture

The prostate carcinoma cell line LNCaP was obtained from ATCC and grown in RPMI 1640 with 10% FCS (Life Technologies, Inc.). Cells were transferred into RPMI-1640 medium with 10% charcoal-stripped fetal calf serum (CS-FCS) 24 h before androgen-regulation experiments. This medium was replaced with fresh CS-FCS media or fresh CS-FCS including 1 nM of the synthetic androgen R1881 (New England Nuclear Life Science Products, Inc.). Cells were harvested for RNA isolation at 0- and 24-h time points.

2.2. Library construction

Total RNA was isolated from androgen-stimulated (LNCaP01) and androgen-starved (LNCaP02) cells using TRIzol (Life Technologies, Inc.) according to the manufacturer's instructions. Poly$(A)^+$ RNA was purified using oligo(dT) chromatography [14]. A unidirectional library was constructed in the pSport1 vector (Life Technologies, Inc.) according to a modification of the Gubler and Hoffman [15] protocol. Poly(A^+) was reverse-transcribed using superscript reverse transcriptase and an oligo(dT) linker/primer containing a *Not*1 site (Life Technologies). Sephacryl-S400 (Pharmacia) was used to size-select the synthesized cDNA and remove excess linkers. Blunt-ended, double-stranded cDNA was ligated with a *Sal*1 adapter, digested with *Not*1, then ligated into *Sal*1–*Not*1 digested pSport1. High-efficiency electrocompetent *Escherichia coli* were transformed using a Bio-Rad GenePulser under recommended conditions. Approximately, 86% of the LNCaP01 and 89% of the LNCaP02 transformants contained inserts. The average insert size for the library was 1.7 kb.

2.3. DNA sequencing

Independent transformant colonies were picked into 100 ul PCR mix [10 mM Tris, pH 8.3, 1.5 mM $MgCl_2$, 50 mM KCl, 120 uM dNTPs, 1 U Taq polymerase (Promega) and 0.12 uM each of VN26 TTTCCCAGTCACGACGTTG-TA and VN27 GTGAGCGGATAACAATTTCAC] and subjected to 40 cycles of 30 s at 94 °C, 30 s at 60 °C and 120 s at 72 °C followed by 10 min at 72° C. Amplified inserts were purified over Sephacryl S-500 (Pharmacia), and 4 ul was used in DNA sequencing reactions using M13 reverse fluorescent-labeled dye primers as detailed in the Prism cycle sequencing kit (Applied Bio-systems, Inc.). Reaction products were electrophoresed on ABI 373 and 377 DNA sequencers.

2.4. Northern analysis

Total RNA was isolated from LNCaP cells using the TRIzol method according to the manufacturer's instructions. Ten micrograms of total RNA was fractionated on 1.2% agarose gels under denaturing conditions and transferred to nylon membrane using the capillary method. Blots were hybridized with cDNA probes labeled with [α-^{32}P]-dCTP using a Random Primers DNA labeling kit (Life Technologies Inc.) according to the manufacturer's protocol. Filters were imaged and quantitated using a phosphor-capture screen and Image Quant software (Molecular Dynamics). β-Actin was used as an internal control for normalizing transcript levels between samples.

2.5. EST assembly, annotation, and comparison

DNA sequences were stored, clustered, and annotated using the PEDB relational database management tools and data analysis pipeline [17].[1] Briefly, vector, *E. coli*, and interspersed repeats were masked in the ESTs using Cross_Match[2] and Repeatmasker.[3] Poor quality sequences, with >50% ambiguous nucleotides ('N') between nucleotides 100 and 500 were discarded. CAP2 [16], a multiple sequence alignment program based on a variant of the Smith–Waterman algorithm, was used to cluster the masked sequence and generate a consensus sequence for each assembly. Each distinct cluster was annotated by searching Unigene,[4] GenBank,[5] and dbEST[6] databases using BLASTN.[7] Annotations were assigned automatically using SmartBlast (Perl 5.0) to select the database match with the lowest *P*-value and the highest BLAST score where the maximum *P*-value was e^{-20} and the minimum BLAST score was 500. Some species required manual reconciliation when either two distinct PEDB species were annotated with the same identification, or when annotations differed between public databases. The Virtual Expression Analysis Tool (VEAT[8]) and scripts written in Perl 5.0 were used for creating transcript species reports. The biological role for each species was assigned using the categories described by Adams et al. [9]. Supplemental information, including a complete list of species and transcript frequencies is available at the PEDB web site. Gene symbols are from the HUGO Gene Nomenclature Committee.

Using statistics described by Audic and Claverie [11], differential gene expression in androgen-stimulated and androgen-deprived cells was inferred based on differential representation of ESTs in cDNA libraries.

2.6. SAGE data acquisition and analysis

The following LNCaP SAGE libraries are listed at the NCBI Library Browser web site[9] and were downloaded from SAGE-map's anonymous FTP site[10]: SAGE_Chen_LNCaP (62,681 tags), SAGE_Chen_LNCaP_no-DHT (65,206 tags), SAGE_CPDR_LNCaP-C (41,848 tags), and SAGE_CPDR_LNCaP-T (44,370 tags). For simplicity, these libraries are hereafter called LNCaP(+)DHT, LNCaP(−)DHT, LNCaP-C and LNCaP-T. Statistical analyses were performed using the software provided at the SAGEmap xProfiler web site.[11]

3. Results

3.1. EST-derived LNCaP transcriptomes

Two cDNA libraries, LNCaP01 and LNCaP02, were constructed from the prostate adenocarcinoma cell line LNCaP under conditions of androgen stimulation and androgen starvation, respectively. Approximately, 2300 ESTs were produced from each library and the sequences were entered into the PEDB [12]. Automated processing of the ESTs to remove short, poor quality, repetitive, and/or vector sequences eliminated 779 ESTs from further analysis. The remaining 4458 ESTs were assembled using the CAP2 sequence assembly program. Each EST cluster was annotated by searching the Unigene, GenBank, and dbEST databases with the CAP2-generated cluster consensus sequences using BLASTN. Clusters annotated with the same database sequence were joined, and all ESTs grouped to the same cluster were assigned the same unique PEDB cluster ID. ESTs for mitochondrial genes were grouped as a single cluster and accounted for approximately 6% of all ESTs. These genes were not further analyzed. In total, 2486 distinct transcript species were identified (Fig. 1): 2240 were homologous to previously identified genes or ESTs, and 252 were not significantly homologous to any public database sequence. The latter species may represent novel genes or previously unsequenced regions of known genes.

The number of distinct transcripts comprising the LNCaP01 and LNCaP02 transcriptomes are quantitatively similar, but qualitatively different. In all, 87% of the species were represented in one transcriptome or the other, but not in both (Fig. 1A). Despite the difference in species composition, the EST frequency distributions of the two samples were similar: nearly 78% of the species are represented by a single EST and only 9% were composed of more than 2 ESTs (Table 1). These distributions are broadly consistent with previous estimates which indicate there are relatively few transcripts expressed in high abundance (5–15 species at 10,000 copies per cell), an intermediate number of moderately abundant transcripts (500 species at 300 copies per cell) and many low abundance transcripts (10,000 different species expressed in 1–15 copies per cell) [17]. In all, 70% of the transcript species with two or more ESTs in either LNCaP01 or LNCaP02 were also present in the other library (Fig. 1B). Thus, while few low abundance transcripts were found in both datasets, most of the high abundance transcripts were found in common.

Functional roles were assigned to each distinct species according to the convention established by Adams et al. [9]. The five primary biological roles were cell division, cell signaling/cell communication, cell structure/motility, cell/organism defense, and metabolism. For graphical presentation, we added the 'androgen-regulated' category to emphasize the primary difference between the experimental samples (Fig. 2). In total, 923 transcript species could be

[1] http://www.pedb.org.
[2] http://www.genome.washington.edu/UWGC/methods.htm.
[3] http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker.
[4] ftp://ncbi.nlm.nih.gov/repository/UniGene/Hs.seq.all.Z.
[5] ftp://ncbi.nlm.nih.gov/blast/db/nt.Z.
[6] ftp://ncbi.nlm.nih.gov/blast/db/est.Z.
[7] http://blast.wustl.edu.
[8] http://www.pedb.org.
[9] http://www.ncbi.nlm.nih.gov/SAGE/sagelb.cgi.
[10] ftp://ncbi.nlm.nih.gov/pub/sage/seq/.
[11] http://www.ncbi.nlm.nih.gov/SAGE/sageexpsetup.cgi.

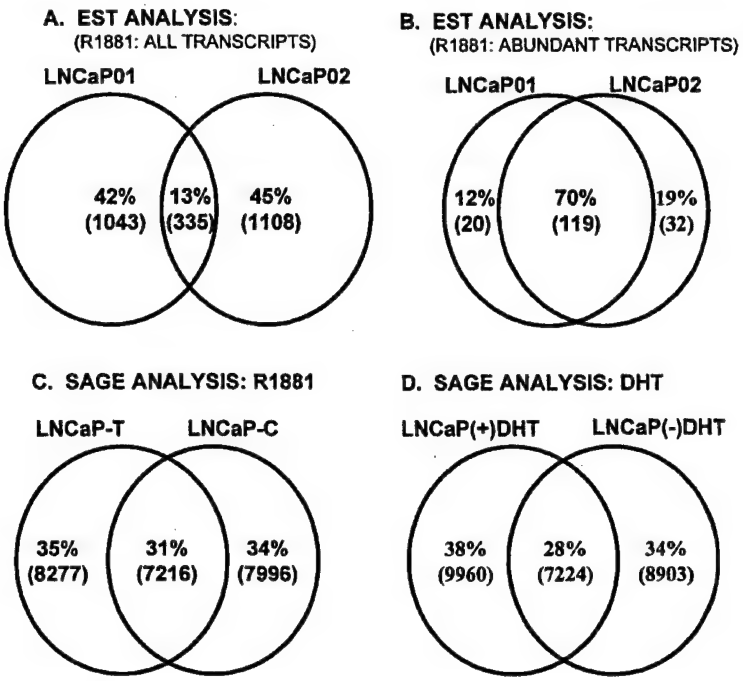


Fig. 1. Summary of LNCaP transcriptome diversity determined by EST and SAGE analysis. Representations of (A) the EST-derived number of all distinct transcripts unique to two LNCaP cell states (synthetic androgen R1881-stimulated LNCaP, LNCaP01; and R1881-starved LNCaP, LNCaP02) and those expressed in common between the two cell states; (B) the EST-derived number of highly and moderately expressed transcripts in LNCaP01 and LNCaP02 (>2 ESTs in one or both libraries) and those expressed in common; (C) SAGE analysis determining the number of distinct transcripts unique and in common between R1881-stimulated and starved LNCaP cells; (D) SAGE analysis determining the number of distinct transcripts unique and in common between DHT-stimulated and starved LNCaP cells.

Table 1
Distribution of molecular species by EST frequency

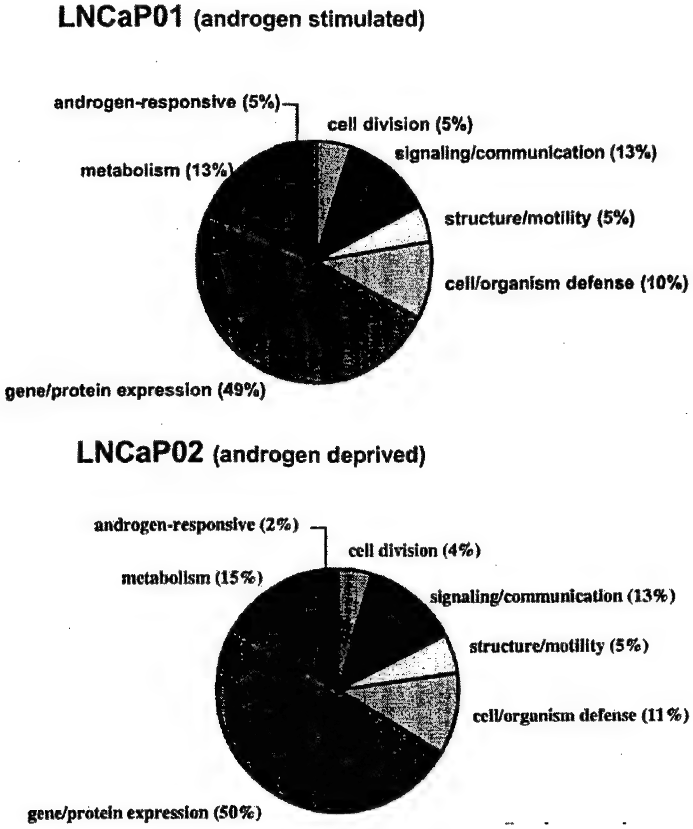
ESTs/species	No. of species (proportion of total)	
	LNCaP01	LNCaP02
1	1064 (0.78)	1133 (0.79)
2	202 (0.15)	199 (0.14)
3	55 (0.04)	56 (0.04)
4	26 (0.02)	23 (0.02)
5	8 (0.01)	8 (0.01)
6	6 (<0.01)	8 (0.01)
>6	17 (0.01)	16 (0.01)
Total	1378	1443

assigned biological roles. A detailed annotation of LNCaP transcripts assigned to these functional roles can be viewed at the PEDB website.[12] Both LNCaP transcript profiles have a similar distribution of species in each functional category (Fig. 2). The protein/gene expression category is the largest, primarily because of the high frequency of ESTs for ribosomal proteins and translation factors. Similar results have been obtained for whole normal prostate tissue [18]. A comparison of the composition of broad functional categories does not reveal a cohort of genes that reflect androgen stimulation or starvation, but differential gene expression in response to androgens is clearly evident for individual genes (Fig. 2). *KLK3/PSA*, an androgen-regulated gene, represents 1.4% of the ESTs in LNCaP01 (derived from androgen-stimulated cells), but only 0.05% of the ESTs in LNCaP02. ESTs for the androgen-response genes *KLK2*, *KLK4*, *ODC1*, *TUBA1*, and *ENO1* were also more abundant in the LNCaP01 library.

3.2. Androgen-regulated genes identified by digital expression analysis

We compared the abundance of each transcript species represented in the androgen-stimulated and androgen-starved transcriptomes using a VEAT [12]. VEAT provides a comprehensive graphical view of transcript frequency, as defined by EST number, between two or more transcriptomes of interest (Fig. 3). Among the species with more than two ESTs in either library, the most extreme difference in EST frequency was observed for *KLK3/PSA*. Twenty-nine *KLK3/PSA* ESTs were isolated from LNCaP01, the library made from androgen-stimulated cells, and only one EST was isolated from LNCaP02 (Table 2). This finding was expected as *KLK3/PSA* is one of the most abundant transcripts

[12] www.pedb.org.


LNCaP01 (androgen stimulated)

LNCaP02 (androgen deprived)

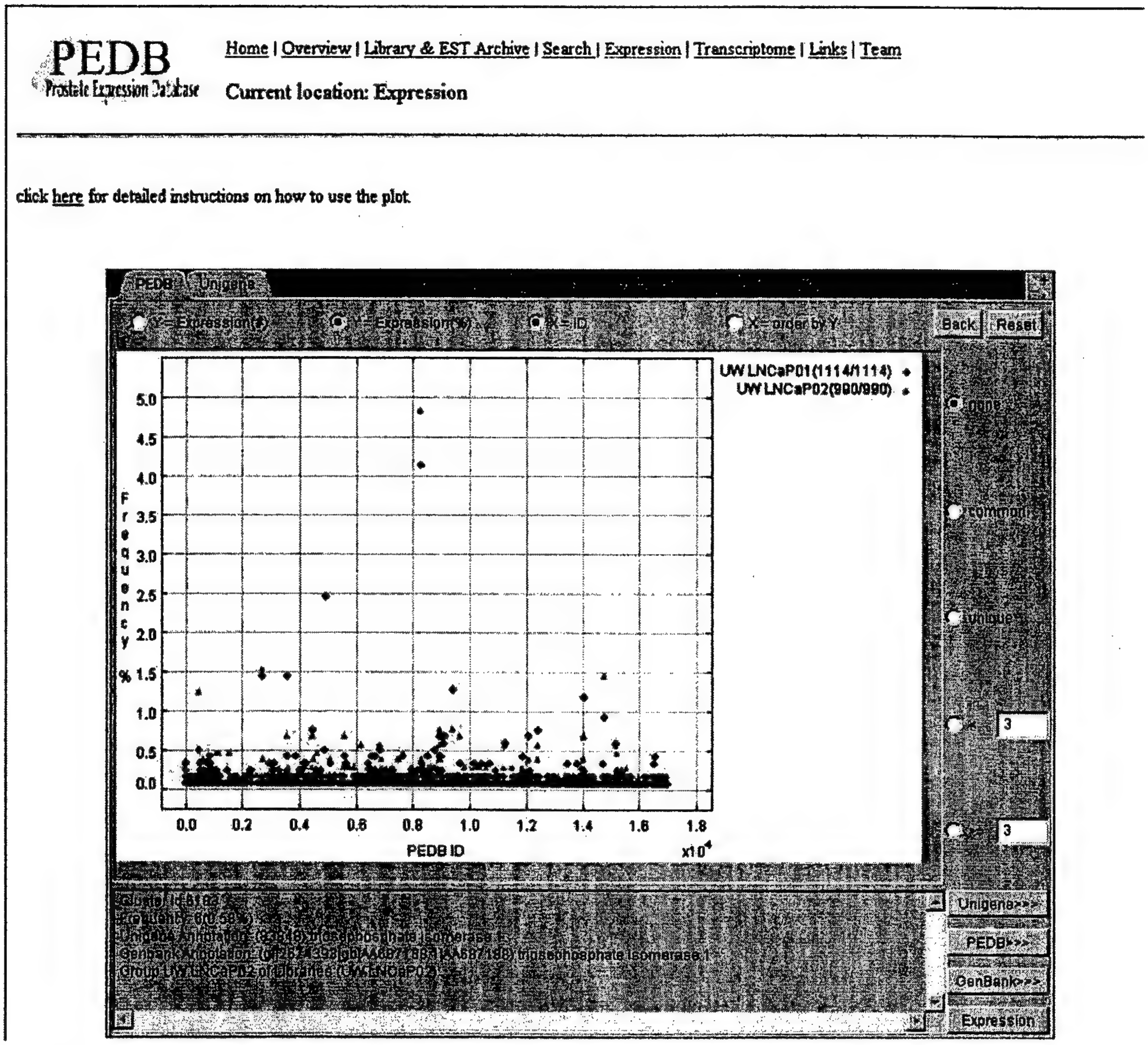
Fig. 2. Functional categorization of the LNCaP cell transcriptome. EST assemblies were annotated against the Genbank and Unigene databases. A putative functional role was assigned based upon categories developed by TIGR (http:www.tigr.org) and the percentage of ESTs corresponding to each role are depicted under cellular conditions of androgen stimulation and androgen starvation.

in the prostate [18] and is known to be transcriptionally regulated by androgens in LNCaP cells.

Additional differences in EST frequencies were seen for many other LNCaP transcripts. Determining the significance of these observations is challenging because of the potential for chance events (e.g. randomly selecting a given cDNA clone from a library) when the event is part of a large population of observable outcomes (e.g. cDNA libraries are complex and comprised of millions of cDNA clones). In order to validate and prioritize more subtle differences in gene expression, we used a statistical approach designed to provide a confidence interval indicating the probability that a given set of observations could occur by chance, or alternatively represents a significant change in expression [11]. Software available on the Internet[13] computes the confidence intervals corresponding to arbitrary significance levels and sample sizes of two datasets N_1 and N_2 [11]. Twenty-one species were predicted to be differentially expressed with a probability exceeding 90%: 9 were increased in response to androgens, and 12 were increased by androgen starvation (Table 2). With the exception of *KLK3/PSA*, none of these genes has previously been reported to be androgen-regulated in the prostate.

To confirm the differential expression statistics, the levels of transcription of *KLK3/PSA* and nine additional genes were examined by Northern analysis (Table 2, Fig. 4). cDNAs representing five different transcripts predicted to be androgen-upregulated by EST analysis were hybridized to Northern blots of RNA extracted from androgen-starved and androgen-stimulated LNCaP cells. Transcipts from each of the five genes were more abundant in androgen-stimulated cells than in androgen-deprived cells. Consistent with the EST frequency data, *KLK3/PSA* expression was increased 35-fold in androgen-stimulated cells compared to androgen-starved cells (Fig. 4). The transcripts encoding keratin 18 (*KRT18*), a gene expressed in prostate secretory cells, were increased 5-fold. FK506 binding protein 5 (*FKBP5*), *DKFZP564K247*, and *UOCRC2* were induced to a lesser extent. In contrast, statistical predictions were inaccurate for four of five putatively down-regulated genes. The steady-state level of DKFZp564K247 RNA was actually increased by androgens, and reduced transcription of

[13] http://igs-server.cnrs-mrs.fr.



PEDB
Prostate Expression Database
Home | Overview | Library & EST Archive | Search | Expression | Transcriptome | Links | Team

Current location: Expression

click here for detailed instructions on how to use the plot

Cluster Id [illegible]
Frequency: 6(0.58%)
Unigene Annotation: (83848) triosephosphate isomerase 1
Genbank Annotation: (gi|2624398|gb|AA687188|AA687188) triosephosphate isomerase 1
Group UW LNCaP02 of libraries (UW LNCaP02)

Fig. 3. Virtual differential expression determined by digital expression profiles. A view of cellular gene expression using the VEAT from the PEDB. Distinct transcripts are assigned a unique database ID and ordered along the *X*-axis. The number of ESTs assembled into each unique transcript (frequency) is displayed on the *Y*-axis as a percentage of the total EST number obtained from each library. Each library is represented by a different symbol (e.g. LNCaP01, triangle; LNCaP02, diamond). Highlighting any data point (using a mouse) provides annotation corresponding to that particular transcript (PEDB reference).

eukaryotic initiation factor 3 subunit 6 (*EIF3S6*), ribosomal protein 27a (*RPS27A*), and basigin (*BSG*) was not confirmed by Northern analysis. Surprisingly, one gene predicted to be decreased by androgen deprivation, the RNA helicase DEAD/H box polypeptide 15 (*DDX15*), was upregulated more than 3-fold by Northern analysis. There are several RNA helicases and our probe may be cross-hybridizing with another closely related androgen-inducible gene. At least, one other androgen-regulated RNA helicase has been reported [19].

In addition to the six androgen-responsive genes identified above, a heat shock protein gene (*HSP90*) was initially identified as androgen-regulated after a preliminary statistical analysis of approximately 1500 LNCaP01 and LNCaP02 ESTs. As the number of ESTs increased, *HSP90* was not differentially expressed based on the arbitrary statistical probability cut-off of $P > 0.90$; however, Northern blot analysis demonstrated a 4-fold increase in *HSP90* expression with androgen stimulation. There are numerous genes in the heat shock protein 90 gene family with strong sequence similarity [20], and our Northern hybridization conditions cannot differentiate between them. Nevertheless, this result confirms that one or more members of the *HSP90* gene family are androgen-responsive.

Table 2
Putative androgen regulated genes in LNCaP01/LNCaP02 libraries ($P \geq 0.9$) and corresponding SAGE data

Gene	ESTs			Androgen Response on Northern blot[a]	SAGE		
	No. of ESTs		Probability of differential expression[b]		SAGE Tag[c]	Probability of differential expression[d,e]	
	LNCaP01[f]	LNCaP02[g]				LNCaP-T/-C[h]	LNCaP(+)DHT/(−)DHT[i]
KLK3/PSA	29	1	$P > 0.99$	+35	GGATGGGGAT	$P = 1.00$ (82/5)	$P = 0.25$ (63/36)
RPLP0	22	9	$0.98 < P < 0.99$	nd[j]	CTCAACATCT	$P = 0.00$ (120/105)	$P = 0.00$ (248/292)
UQCRC2	5	0	$0.96 < P < 0.97$	+1.3	AAAGTCAGAA	$P = 0.16$ (6/8)	$P = 0.16$ (6/5)
FKBP5	4	0	$0.93 < P < 0.94$	+1.9	GTTCCAGTGA	$P = 0.66$ (6/0)	$P = 0.39$ (0/2)
DKFZP564K247	4	0	$0.93 < P < 0.94$	+1.7	TATCGGGAAT	–	$P = 0.29$ (2/1)
PHGDH	4	0	$0.93 < P < 0.94$	nd	TTACCTCCTT	$P = 0.22$ (22/12)	$P = 0.15$ (65/40)
KRT18	4	0	$0.93 < P < 0.94$	+5.0	CAAACCATCC	$P = 0.12$ (22/14)	$P = 0.02$ (27/35)
RPS25	6	1	$0.93 < P < 0.94$	nd	AATAGGTCCA	$P = 0.00$ (53/51)	$P = 0.06$ (132/84)
SFTPD	9	3	$0.90 < P < 0.91$	nd	–	–	–
E1F3S6	0	6	$0.98 < P < 0.99$	+1.2	AATATTGAGA	$P = 0.07$ (11/10)	$P = 0.33$ (12/6)
FTL	0	5	$0.96 < P < 0.97$	nd	CCCTGGGTTC	$P = 0.24$ (9/15)	$P = 0.15$ (22/37)
DDX15	0	4	$0.93 < P < 0.94$	+3.5	ATCGTTGTAA	$P = 0.37$ (4/1)	$P = 0.47$ (3/0)
RPS27A	0	4	$0.93 < P < 0.94$	+1.3	AACTAACAAA	$P = 0.15$ (16/10)	$P = 0.14$ (49/31)
ACADVL	0	4	$0.93 < P < 0.94$	nd	GCCGCCCTGC	$P = 0.13$ (6/6)	$P = 0.48$ (8/20)
KIAA0101	0	4	$0.93 < P < 0.94$	nd	ATGATTTATT	$P = 0.21$ (3/4)	$P = 0.47$ (3/0)
DKFZp564D0462	0	4	$0.93 < P < 0.94$	−2.6	CAGTTCTCAC	$P = 0.29$ (1/1)	$P = 0.40$ (2/0)
RPS15A	0	4	$0.93 < P < 0.94$	nd	GACAAAAAAA	$P = 0.26$ (27/14)	$P = 0.18$ (12/8)
RPS15A	–	–	–	–	GACTCTGGTG	$P = 0.16$ (11/7)	$P = 0.00$ (36/41)
DED	0	4	$0.93 < P < 0.94$	nd	GCACCTATTG	$P = 0.29$ (2/1)	$P = 0.35$ (0/1)
Species1145	0	4	$0.93 < P < 0.94$	nd	–	–	–
BSG	1	6	$0.92 < P < 0.93$	−1.02	GCCGGGTGGG	$P = 0.06$ (11/11)	$P = 0.00$ (216/341)
TPI1	1	6	$0.92 < P < 0.93$	nd	TGAGGGAATA	$P = 0.01$ (33/29)	$P = 0.02$ (39/32)

[a] Ratio of normalized signal intensity from RNA of hormone stimulated/starved cells.
[b] [11].
[c] Most abundant unique tag.
[d] [35].
[e] Tag frequency in hormone stimulated/starved samples.
[f] 2222 ESTs.
[g] 2236 ESTs.
[h] ~42,000 tags per library.
[i] ~62, 000 tags per library.
[j] nd, not done.

3.3. Comparison of EST and SAGE digital expression profiles

An alternate method of acquiring qualitative and quantitative transcript profiles is by the SAGE. Rather than producing gene tags of 300–500 nucleotides, the SAGE method generates sequence tags of approximately 10 nucleotides in length. This difference allows 10–30-fold more SAGE tags to be acquired per sequencing reaction, thus, deeper transcript profiles can be obtained more efficiently. However, the short tag length may introduce ambiguity when assigning a tag to a specific gene [21].

Data from two independent SAGE profiling experiments examining androgen-regulated gene expression in LNCaP cells were obtained from the SAGEmap website at NCBI.[14] Descriptions of the libraries indicated that one SAGE dataset, designated LNCaP(−)DHT/(+)DHT, was derived from LNCaP cells grown in hormone-depleted media for 3 months (LNCaP(−)DHT) and then stimulated with 1 nM DHT (LNCaP(+)DHT) for 24 h. Approximately 63,000 tags were sequenced from each library. The second SAGE dataset, LNCaP-T/-C, was derived from cells grown in hormone-depleted media for 5 days (LNCaP-C), then stimulated with 10^{-8} M R1881 for 24 h (LNCaP-T). Approximately, 42,000 tags were sequenced from each library. The distribution of expressed genes in each pair of SAGE libraries is given in Fig. 1B and C.

Theoretical and empirical data suggest that roughly 650,000 transcripts must be sampled to identify all but very rare mRNAs in the cell [22]. Thus, neither our study nor the SAGE datasets were large enough to thoroughly sample transcript diversity in the LNCaP cells, and neither dataset is capable of identifying differential gene expression among low abundance transcripts. Broadly, genes with a role in protein synthesis (ribosomal proteins and translation initiation factors) were the most abundant transcripts in both our EST data and the SAGE profiles. Interestingly, the EST approach identified approximately 200 transcript species

[14] http://www.ncbi.nlm.nih.gov/SAGE/.

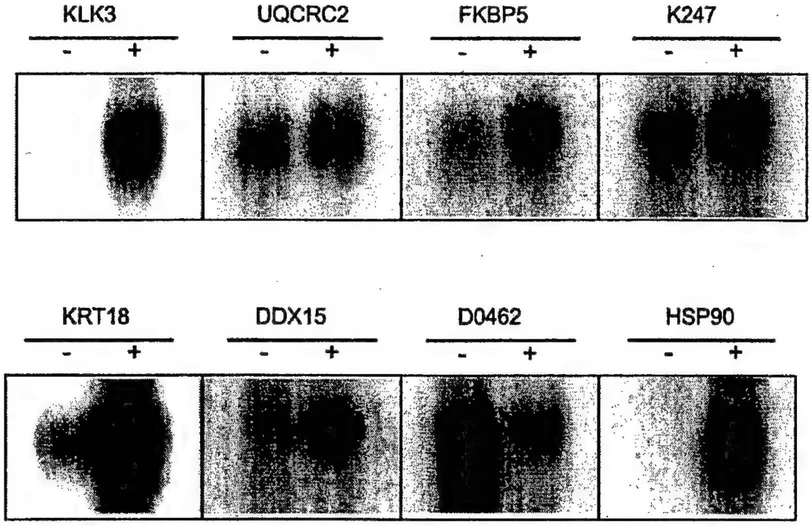


Fig. 4. Northern blots of eight androgen regulated genes predicted to be differentially expressed by virtual EST analysis. K247 is *DKFZP564K247* and D0462 is *DKFZP564D0462*. 'Minus', total RNA from androgen-starved LNCaP cells. 'Plus', total RNA from LNCaP cells treated with 1 nM R1881.

with corresponding Unigene entries that were not observed in the SAGE libraries. Conversely, the SAGE studies identified hundreds of transcripts that were not observed in the EST assemblies. Thus, these studies complement each other in creating an inventory representing the LNCaP cell transcriptome.

Transcripts with a high probability of differential expression between each pair of SAGE profiles were identified using the SAGEmap xProfiler. Despite a 10-fold difference in sample size, the SAGE and EST studies identified similar numbers of putative androgen-responsive genes (cut-off $P = 0.9$). In the EST analysis, 21 genes had a high probability

Table 3
Known androgen-response genes exhibiting differential expression in one or more libraries ($P \geq 0.6$)

Gene	ESTs			SAGE tag[a]	SAGE		
	No. of ESTs		Probability of differential expression[b]		Probability of differential expression[c,d]		Prostate-enriched[e]
	LNCaP01[f]	LNCaP02[g]			LNCaP-T/-C[h]	LNCaP(+)DHT/(−) DHT[i]	
CLTB	0	0	$0.00 < P < 0.10$	GGCTGGGCCT	$P = 0.45$ (3/0)	$P = 0.73$ (2112)	−
DBI	1	0	$0.50 < P < 0.60$	TGTTTATCCT	$P = 0.77$ (13/2)	$P = 0.03$ (20/18)	−
ENO1	6	3	$0.60 < P < 0.70$	GTGTCTCATC	$P = 0.13$ (9/12)	$P = 0.04$ (15/14)	−
KLK2	3	0	$0.80 < P < 0.90$	CTGTGGTTTA	$P = 0.39$ (2/0)	$P = 0.80$ (810)	+
	–	–	–	CTGTGGTTAA	–	$P = 0.76$ (14/3)	+
KLK3	29	1	$P > 0.99$	GGATGGGGAT	$P = 1.00$ (82/5)	$P = 0.25$ (63/36)	+
KLK4	2	0	$0.70 < P < 0.80$	AAATTGACCC	$P = 0.35$ (1/0)	$P = 0.51$ (2/8)	+
ODC1	4	1	$0.70 < P < 0.80$	TGCGTGGTCA	$P = 0.35$ (1/0)	–	−
	–	–	–	ATGCAGCCAT	–	$P = 0.11$ (7/7)	−
PDHA1	0	0	$0.00 < P < 0.10$	CAGTTTGTAC	$P = 0.60$ (5/0)	$P = 0.28$ (4/2)	−
PMEPA1[j]	1	0	$0.50 < P < 0.60$	TGATGTCTGG	$P = 1.00$ (29/1)	$P = 0.47$ (7/2)	+
TUBA1	4	1	$0.70 < P < 0.80$	GAGGAGGGTG	$P = 0.29$ (2/4)	$P = 0.44$ (5/13)	−
UGT2B17	0	0	$000 < P < 0.10$	GAGGGTTTTA	$P = 0.62$ (0/5)	$P = 0.40$ (4/1)	−
VEGF	1	1	$0.00 < P < 0.10$	TTTCCAATCT	$P = 0.29$ (1/2)	$P = 0.69$ (610)	−

[a] Most abundant unique tag.
[b] [11].
[c] [35].
[d] Tag frequency in hormone stimulated/starved cells.
[e] More abundant in the prostate than in most other tissues.
[f] 222 ESTs.
[g] 2236 ESTs.
[h] ~42,000 tags per library.
[i] ~62,000 tags per library.
[j] Tag inferred from [34].

of differential expression (9 up-regulated, 12 down-regulated) while 17 unique tags were identified in the SAGE LNCaP-T/-C study (6 up-regulated, 11 down-regulated), and 23 were identified in the SAGE LNCAP(+)DHT/(−)DHT study (17 up-regulated, 6 down-regulated). Surprisingly, with the exception of *KLK3/PSA*, all of the identified genes were different across the three datasets. *KLK3/PSA* had a high probability of differential expression in both our EST dataset ($P > 0.99$) and the LNCAP-T/-C dataset ($P = 1.0$). The only other potential androgen-regulated gene in the EST data that had a moderate probability of differential expression based on SAGE was FK506 binding protein 5 (*FKBP5*; $P = 0.66$, LNCaP-T/-C). The three genes that we confirmed to be differentially expressed by Northern blot analysis (keratin 18, 3-phosphoglycerate dehydrogenase, and *DKFZP564K247*) were not expressed at significantly different levels ($P < 0.30$) in the two SAGE datasets.

A review of published literature identified 75 genes reported to be androgen-responsive in one or more human tissues (see PEDB [15]). Twenty-three of these genes had corresponding EST tags; 47 had LNCaP-T/-C SAGE tags; and 55 had LNCaP(+)DHT/(−)DHT SAGE tags. Thus, SAGE sampling of 10-fold more transcripts only doubled the number of observed, previously-described, androgen-regulated genes. The genes identified in the EST dataset are not just a subset of those found in the larger SAGE datasets: *TMPRSS2*, a serine protease gene whose transcription is stimulated by androgen in LNCaP cells [23], was represented in the EST data, but not in the SAGE libraries. Only 12 of the 75 known androgen-response genes had even a moderate probability of differential expression ($P \geq 0.6$) in one or more datasets (Table 3), and there is no case where statistical predictions agree across all three data sets. Six of the twelve genes were predicted to be androgen inducible in the EST dataset, compared to five genes in the LNCaP-T/-C dataset and three in the LNCaP(+)DHT/(−)DHT dataset. The two SAGE studies, with similar numbers of tags, predicted completely different cohorts of up-regulated genes (Table 3).

4. Discussion

The identification and quantitation of the complement of genes expressed in a cell or tissue provides a framework for understanding biological properties and establishes a tool set for functional studies. Several methods have been developed for the comprehensive analysis of gene expression in complex biological systems. We have investigated the application of two procedures, EST profiling and SAGE, to characterize the transcriptome of prostate adenocarcinoma cells and to identify the cohort of genes regulated directly or indirectly by androgenic hormones. The EST profiles obtained from two LNCaP cDNA libraries identified 2486 distinct transcripts. Of these, 336 were expressed in common. The total number of transcripts, we identified in this study represents about 12–17% of the total complexity found in prostate epithelium [24] and likely includes all highly expressed, many moderately expressed and relatively few rarely expressed transcripts. Many of these genes were previously identified in other tissues, but were not known to be expressed in the prostate. In all, 252 new transcripts were identified that are not represented in any public database. Since over 2.2 million human ESTs are present in dbEST (release 081800), some of the unknown transcripts may be exclusively expressed in the prostate epithelium. These findings support the continued utility of cataloging transcripts from specialized tissue sources. These newly identified cDNAs can be tested for tissue-specific expression and can be used both to facilitate the identification of exons in the context of the human genome project and to enhance the positional cloning of prostate cancer susceptibility genes.

Androgens regulate numerous processes in prostate epithelial cells that include cell division, cell quiescence, apoptosis, lipid metabolism, and the production of specialized secretory proteins such as KLK3/PSA. Of the 2486 distinct transcripts identified in the LNCaP transcriptome, 364 (14%) showed at least a 2-fold difference in expression following exposure to androgens. Statistical analysis reduced this number to 21 genes with a high probability of differential expression ($P \geq 0.9$). Ten were further tested by Northern analysis which confirmed six were indeed transcriptionally regulated by androgen; *KLK3/PSA*, *FKBP5*, *KRT18*, *DDX15*, and *DKFZP564D0462*. In addition, *HSP90* was identified as an androgen-response gene by Northern blot analysis. These data identify five genes as new members of the androgen-response network, since only *KLK3/PSA* was previously known to be androgen-responsive. The lack of complete concordance between the digital expression results and Northern analysis can be partly explained by cross-hybridization to highly-homologous gene family members, alternative splicing events, and the lack of Northern sensitivity to alterations in low abundance transcripts.

The genes found in this study to be transcriptionally sensitive to androgen have diverse functions. *KLK3/PSA* is a highly abundant serine protease with known androgen-response elements in the promoter region [25] and prostate-enriched expression. Keratin 18 is a marker for prostate luminal cells [26] but is found in a variety of epithelia. The *DKFZP564D0462* gene encodes a putative seven transmembrane-domain protein that is expressed in a variety of tissues. The DEAD/H box polypeptide 15 gene is a putative RNA helicase similar to a yeast gene required for mRNA splicing [27]. Another RNA helicase, GRTH, is up-regulated in testis in response to androgen [19]. These genes may play a role in steroidogenesis or androgen-mediated stimulation of protein synthesis. *HSP90* binds and activates the androgen receptor. *FKBP5*, another gene predicted to be up-regulated in LNCaP cells, interacts with *HSP90* in func-

[15] http://www.pedb.org.

tionally mature progesterone complexes [28]. Hence, both *HSP90* and *FKBP5* may be up-regulated to facilitate signal transduction through the androgen receptor.

While general trends in gene expression were similar with respect to the overall effects of androgens, why was little concordance found between EST data and the SAGE data in terms of the expression of specific genes? In part this may be attributable to relatively small overall sample sizes and the limitations of statistical confidence. Cloning or sequencing biases could be unequally introduced by the experimental approaches, and ambiguity in SAGE tag assignment may affect a subset of genes. However, an alternative explanation is that each method accurately reflects the state of cellular gene expression, and the differences are attributable to the actual in vitro conditions. There will be some variation in transcript levels even under optimal conditions that may relate to cell density, growth media, and other factors. At present, we do not know the precise effects of protracted androgen starvation on LNCaP cells, but the extended starvation of cells used to create the LNCaP(+)DHT/(−)DHT libraries (3 months), could have selected for altered gene expression. In this regard, it is noteworthy that *KLK3/PSA*, one of the most abundant androgen regulated genes, was not differentially expressed in the LNCaP(+)DHT/(−)DHT dataset (Table 3). Cell-line history may also affect transcription. LNCaP may have undergone significant physiological adaptation and genomic change during maintenance in different laboratories. Esquenet et al. [29] observed a marked decrease in the ability of androgen to induce *KLK3/PSA* transcription in LNCaP cells of high passage number relative to cells of low passage number. And LNCaP cells can undergo "proliferative shut-off" in response to androgen [30]. These experimental differences may be analogous to the heterogeneity observed between individual cancers and may be reflected in the cellular transcriptomes assayed by digital-expression profiles.

Another intriguing possibility is that different androgens and androgen concentrations activate or repress subnetworks of the androgen-response program. Testosterone, DHT, and synthetic androgens such as R1881 induce a concentration-dependent biphasic growth response in LNCaP cells that may be influenced by the relative activities of growth-promoting and growth-suppressing genes [31]. Different ligands or ligand concentrations may recruit distinct AR co-activator molecules that dictate the subset of genes to be activated [32,33]. Of interest, a report describing the cloning and characterization of the gene corresponding to the SAGE tag exhibiting the greatest androgen-induction (29-fold) in the LNCaP(+)DHT/(−)DHT SAGE dataset was recently published [34]. By Northern analysis, the expression of this gene, *PMEPA1*, was shown to increase only 2-fold with 10^{-10} M R1881, but nearly 5-fold with 10^{-8} M R1881; the concentration used in the SAGE experiments. The 10^{-9} M R1881 concentration used in our EST experiments did not induce a detectable increase in *PMEPA1* EST frequency.

At present, financial and technological barriers make it impractical to simultaneously test all known genes for expression in the prostate. Inventories of genes from cell lines such as LNCaP, which are used extensively as model systems for studying prostate cancer, can help alleviate this problem by identifying the subset of genes of relevant to the biological system under study. Additional SAGE and EST data are needed to identify rare transcripts and to increase statistical power required for robust digital expression studies. In addition to their demonstrated utilities as gene discovery and analysis tools, the digital expression profiling methods used here can also greatly facilitate the construction of microarray-based reagents suitable for applications where higher throughput is required.

Acknowledgements

We thank Barbara Trask for critical review of the manuscript, David Haynor and Roger Bumgarner for helpful discussions, and Sue Heiner for manuscript preparation. This work is supported in part by grants (DAMD 17-98-1-8499 and PC991274) from the Department of Defense, a Grant (CA75173) from the National Cancer Institute, and the CaPCURE foundation.

References

[1] A.O. Brinkmann, L.J. Blok, P.E. de Ruiter, P. Doesburg, K. Steketee, C.A. Berrevoets, J. Trapman, Mechanisms of androgen receptor activation and function, J. Steroid Biochem. Mol. Biol. 69 (1999) 307–313.

[2] J. Trapman, K.B. Cleutjens, Androgen-regulated gene expression in prostate cancer, Seminars Cancer Biol. 8 (1997) 29–36.

[3] B. Ewing, P. Green, Analysis of expressed sequence tags indicates 35,000 human genes, Nat. Genet. 25 (2000) 232–234.

[4] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, et al., The sequence of the human genome, Science 291 (2001) 1304–1351.

[5] V.E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M.A. Basrai, D.E. Bassett, P. Hieter, B. Vogelstein, K.W. Kinzler, Characterization of the yeast transcriptome, Cell 88 (1997) 243–251.

[6] G.G. Lennon, H. Lehrach, Hybridization analyses of arrayed cDNA libraries, Trends Genet. 7 (1991) 314–317.

[7] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science 270 (1995) 467–470.

[8] L. Wodicka, H. Dong, M. Mittmann, M.-H. Ho, D.J. Lockhart, Genome-wide expression monitoring of *Saccharomyces cerevisae*, Nature Biotechnol. 15 (1997) 1359–1367.

[9] M.D. Adams, A.R. Kerlavage, R.D. Fleischman, R.A. Fuldner, C.J. Bult, N.H. Lee, E.F. Kirkness, K.G. Weinstock, J.D. Gocayne, O. White, Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence, Nature 377 (Suppl. 28) (1995) 3–174.

[10] V.E. Velculescu, L. Zhang, B. Vogelstein, K.W. Kinzler, Serial analysis of gene expression, Science 270 (1995) 384–387.

[11] S. Audic, J.M. Claverie, The significance of digital gene expression profiles, Genome Res. 7 (1995) 986–995.

[12] V. Hawkins, D. Doll, R. Bumgarner, T. Smith, C. Abajian, L. Hood, P.S. Nelson, PEDB: the prostate expression database, Nucl. Acids Res. 27 (1999) 204–208.

[13] G.J. van Steenbrugge, M. Groen, J.W. van Dongen, J. Bolt, H. van der Korput, J. Trapman, M. Hasenson, J. Horoszewicz, The human prostatic carcinoma cell line LNCaP and its derivatives: an overview, Urol. Res. 17 (1959) 71–77.

[14] T. Maniatis, E.F. Fritsch, J. Sambrook, Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1982.

[15] U. Gubler, B.J. Hoffman, A simple and very efficient method for generating cDNA libraries, Gene 25 (1983) 263–269.

[16] X. Huang, An improved sequence assembly program, Genomics 33 (1996) 21–31.

[17] N.D. Hastie, J.O. Bishop, The expression of three abundance classes of messenger RNA in mouse tissue, Cell 9 (1976) 761–774.

[18] P.S. Nelson, W.L. Ng, M. Schummer, L.D. True, A.Y. Liu, R.E. Bumgarner, C. Ferguson, A. Dimak, L. Hood, An expressed-sequence-tag database of the human prostate: sequence analysis of 1168 cDNA clones, Genomics 47 (1998) 12–25.

[19] P.Z. Tang, C.H. Tsai-Morris, M.L. Dufau, A novel gonadotropin-regulated testicular RNA helicase: a new member of the dead-box family, J. Biol. Chem. 274 (1999) 37932–37940.

[20] K. Ozawa, Y. Murakami, T. Eki, E. Soeda, K. Yokoyama, Mapping of the gene family for human heat-shock protein 90 alpha to chromosomes 1,4,11, and 14, Genomics 12 (1992) 214–220.

[21] J. Stollberg, J. Urschitz, Z. Urban, C.D. Boyd, A quantitative evaluation of SAGE, Genome Res. 10 (2000) 1241–1248.

[22] V.E. Velculescu, S.L. Madden, L. Zhang, A.E. Lash, J. Yu, C. Rago, A. Lal, C.J. Wang, G.A. Beaudry, K.M. Ciriello, B.P. Cook, M.R. Dufault, A.T. Ferguson, Y. Gao, T.C. He, H. Hermeking, S.K. Hiraldo, P.M. Hwang, M.A. Lopez, H.F. Luderer, B. Mathews, J.M. Petroziello, K. Polyak, L. Zawel, K.W. Kinzler, Analysis of human transcriptomes, Nat. Genet. 23 (1999) 387–388.

[23] B. Lin, C. Ferguson, J.T. White, S. Wang, R. Vessella, L.D. True, L. Hood, P.S. Nelson, Prostate-localized and androgen-regulated expression of the membrane-bound serine protease TMPRSS2, Cancer Res. 59 (1999) 4180–4184.

[24] L. Zhang, W. Zhou, V.E. Velculescu, S.E. Kern, R.H. Hruban, S.R. Hamilton, B. Vo-gelstein, K.W. Kinzler, Gene expression profiles in normal and cancer cells, Science 276 (1997) 1268–1272.

[25] K.B. Cleutjens, C.C. van Eekelen, H.A. van der Korput, A.O. Brinkmann, J. Trap-man, Two androgen response regions cooperate in steroid hormone regulated activity of the prostate-specific antigen promoter, J. Biol. Chem. 271 (1996) 6379–6388.

[26] E.R. Sherwood, L.A. Berg, N.J. Mitchell, J.E. McNeal, J.M. Kozlowski, C. Lee, Differential cytokeratin expression in normal, hyperplastic and malignant epithelial cells from human prostate, J. Urol. 143 (1990) 167–171.

[27] O. Imamura, M. Sugawara, Y. Furuichi, Cloning and characterization of a putative human RNA helicase gene of the DEAH-box protein family, Biochem. Biophys. Res. Commun. 240 (1997) 335–340.

[28] S.C. Nair, R.A. Rimerman, E.J. Toran, S. Chen, V. Prapapanich, R.N. Butts, D.F. Smith, Molecular cloning of human *FKBP51* and comparisons of immunophilin interactions with *Hsp90* and progesterone receptor, Mol. Cell. Biol. 17 (1997) 594–603.

[29] M. Esquenet, J.V. Swinnen, W. Heyns, G. Verhoeven, LNCaP prostatic adenocarcinoma cells derived from low and high passage numbers display divergent responses not only to androgens but also to retinoids, J. Steroid Biochem. Mol. Biol. 62 (1997) 391–399.

[30] P. Geck, J. Szelei, J. Jimenez, T.M. Lin, C. Sonnenschein, A.M. Soto, Expression of novel genes linked to the androgen-induced, proliferative shutoff in prostate cancer cells, J. Steroid Biochem. Mol. Biol. 63 (1997) 211–218.

[31] E.G. Langeler, C.J. van Uffelen, M.A. Blankenstein, G.J. van Steenbrugge, E. Mulder, Effect of culture conditions on androgen sensitivity of the human prostatic cancer cell line LNCaP, Prostate 23 (1993) 213–223.

[32] P.W. Hsiao, T.H. Thin, L.D. Lin, C. Chang, Differential regulation of testosterone vs. Salpha-dihydrotestosterone by selective androgen response elements, Mol. Cell. Biochem. 206 (2000) 169–175.

[33] S. Yeh, H.C. Chang, H. Miyamoto, H. Takatera, M. Rahman, H.Y. Kang, T.H. Thin, H.K. Lin, C. Chang, Differential induction of the androgen receptor transcriptional activity by selective androgen receptor coactivators, Keio. J. Med. 48 (1999) 87–92.

[34] L.L. Xu, N. Shanmugam, T. Segawa, I.A. Sesterhenn, D.G. McLeod, J.W. Moul, S. Srivastava, A novel androgen-regulated gene, *PMEPA1*, located on chromosome 20q13 exhibits high level expression in prostate, Genomics 66 (2000) 257–263.

[35] A.E. Lash, C.M. Tolstoshev, L. Wagner, G.D. Schuler, R.L. Strausberg, G.J. Rig-gins, S.F. Altschul, SAGEmap: a public gene expression resource, Genome Res. 10 (2000) 1051–1060.

218–220 *Nucleic Acids Research, 2002, Vol. 30, No. 1* © 2002 *Oxford University Press*

The human (PEDB) and mouse (mPEDB) Prostate Expression Databases

Peter S. Nelson[1,2,*], Colin Pritchard[1], Denise Abbott[1] and Nigel Clegg[1]

[1]Division of Human Biology and [2]Division of Clinical Research, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109-1024, USA

Received September 17, 2001; Accepted September 18, 2001

ABSTRACT

The Prostate Expression Databases (PEDB and mPEDB) are online resources designed to allow researchers to access and analyze gene expression information derived from the human and murine prostate, respectively. Human PEDB archives more than 84 000 Expressed Sequence Tags (ESTs) from 38 prostate cDNA libraries in a curated relational database that provides detailed library information including tissue source, library construction methods, sequence diversity and sequence abundance. The differential expression of each EST species can be viewed across all libraries using a Virtual Expression Analysis Tool (VEAT), a graphical user interface written in Java for intra- and inter-library sequence comparisons. Recent enhancements to PEDB include (i) the development of a murine prostate expression database, mPEDB, that complements the human gene expression information in PEDB, (ii) the assembly of a non-redundant sequence set or 'prostate unigene' that represents the diversity of gene expression in the prostate, and (iii) an expanded search tool that supports both text-based and BLAST queries. PEDB and mPEDB are accessible via the World Wide Web at http://www.pedb.org and http://www.mpedb.org.

INTRODUCTION

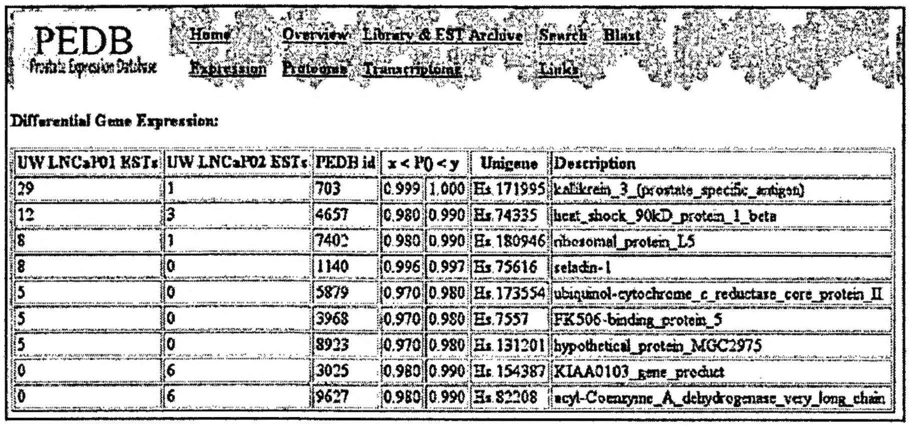
Diseases of the prostate are among the most common pathologies to afflict aging men. Prostate carcinoma is the most frequently diagnosed non-cutaneous malignancy in the US with more than 180 000 new cases estimated for 2001 (1). In order to characterize molecular alterations that accompany prostate disease processes and provide resources for virtual and physical analyses, we have developed the Prostate Expression Database (PEDB) (2). PEDB serves as a centralized collection of gene expression information derived from the human prostate that is organized in a fashion suitable for sequence-based queries, assessment of gene expression diversity, and comparative expression analyses. Expressed Sequence Tags (ESTs) and full-length cDNA sequences derived from 38 human prostate cDNA libraries are archived and represent gene expression profiles reflecting a wide spectrum of normal, benign and malignant prostate disease states. Detailed library information including tissue source, library construction methods, sequence diversity and sequence abundance are maintained in a relational database management system (RDBMS). Prostate ESTs are assembled into distinct species groups using the sequence assembly program Phrap, and annotated with information from the GenBank, dbEST and Unigene public sequence databases.

In recognition of the emerging uses of the mouse as a model system for the study of normal and pathological prostate development, we have developed a database complementary to PEDB that serves to archive and analyze murine prostate gene expression information. The mouse Prostate Expression Database (mPEDB) currently comprises >6000 ESTs from five mouse prostate cDNA libraries constructed from distinct developmental stages and anatomical locations. A detailed description of the database development, data inventory and utilities is available online: www.pedb.org/OVERVIEW/ (Table 1).

Table 1. Table of contents for PEDB overview (http://www.pedb.org/OVERVIEW)

1.	Introduction to PEDB
2.	PEDB Information
	2.a. Construction
	2.b. Dataflow
	2.c. Build Process
	2.d. Current PEDB Build
3.	PEDB Utilities
	3.a. BLAST Queries
	3.b. Search Engine
	3.c. Prostate Unigene
	3.d. Virtual Analysis Expression Tool (VEAT)
	3.e. Prostate Transcriptome
	3.f. Prostate Proteome
4.	mPEDB
5.	PEDB References and Resources

*To whom correspondence should be addressed at: Division of Human Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, PO Box 19024, Seattle, WA 98109-1024, USA. Tel: +1 206 667 3377; Fax: +1 206 667 2917; Email: pnelson@fhcrc.org



Differential Gene Expression:

UW LNCaP01 ESTs	UW LNCaP02 ESTs	PEDB id	x < P() < y		Unigene	Description
29	1	703	0.999	1.000	Hs.171995	kallikrein_3_(prostate_specific_antigen)
12	3	4657	0.980	0.990	Hs.74335	heat_shock_90kD_protein_1_beta
8	1	7402	0.980	0.990	Hs.180946	ribosomal_protein_L5
8	0	1140	0.996	0.997	Hs.75616	seladin-1
5	0	5879	0.970	0.980	Hs.173554	ubiquinol-cytochrome_c_reductase_core_protein_II
5	0	3968	0.970	0.980	Hs.7557	FK506-binding_protein_5
5	0	8923	0.970	0.980	Hs.131201	hypothetical_protein_MGC2975
0	6	3025	0.980	0.990	Hs.154387	KIAA0103_gene_product
0	6	9627	0.980	0.990	Hs.82208	acyl-Coenzyme_A_dehydrogenase_very_long_chain

Figure 1. Output of differential expression analysis with statistical filtering. The annotated ESTs in two prostate cDNA libraries were compared for relative abundance levels. The output of the analysis provides (i) the number of ESTs in each library corresponding to a specific transcript, (ii) the PEDB identification number, (iii) the statistical probability, P, of differential expression between the two library datasets, (iv) the Unigene database accession number, and (v) a description of the gene based upon GenBank or Unigene annotation.

PEDB DATA AND ANALYSIS TOOLS

PEDB consists of archives of ESTs derived from 38 human prostate cDNA libraries. These ESTs are obtained from public sequence repositories such as GenBank (3), the database of ESTs (dbEST) (4), the Cancer Genome Anatomy Project (CGAP) (5), The Institute for Genome Research (TIGR) or from in-house EST sequencing projects. Sequence processing and curation involves a pipeline of sequence submission, sequence masking, sequence assembly and assembly annotation that now incorporates quality-based assemblies using Phred and Phrap base-calling and sequence assembly algorithms (6,7) (www.pedb.org/OVERVIEW/). Assembled consensus sequences are used for BLAST queries against the Unigene, GenBank and dbEST databases to provide cluster annotation and to further facilitate the assembly process.

The most recent build of PEDB ESTs was assembled starting with 84 832 prostate ESTs. Portions of EST sequences with homology to cloning vectors, *Escherichia coli* genomic DNA and human repetitive DNA sequences were masked. Sequences annotating to the mitochondrial genome were removed and the remaining ESTs with >300 bp of high quality sequence were admitted to the assembly process. A total of 68 426 high-quality ESTs were assembled using Phrap to produce 28 182 clusters. Each cluster was annotated by searching the Unigene, GenBank and dbEST databases using BLASTN. Clusters annotating to the same database sequence were joined to further reduce the number of distinct clusters to 20 187. These annotated assemblies represent the prostate transcriptome: that portion of the genome that is used or expressed in the prostate.

The primary work sites of PEDB involve text-based queries and a BLAST interface for sequence-based searches against PEDB and Unigene datasets. Dynamic gene expression profiles based upon EST assembly and annotation information can be generated using the Virtual Expression Analysis Tool (VEAT). The VEAT provides user-directed inter- and intra-library analysis of transcript abundance, diversity and differential expression. We have recently incorporated a statistical algorithm developed by Audic and Claverie (8) that can determine probabilities of differential transcript abundance levels in datasets comprised of varying numbers of sequences. We have used these tools to identify prostate genes regulated by androgens and genes differentially expressed between adenocarcinoma and small cell carcinoma of the prostate (Fig. 1).

MOUSE PEDB (mPEDB)

The mouse represents a versatile model organism for studying development, genetics, behavior and disease. Several murine models of prostate carcinogenesis have recently been reported (9,10), and the mouse has been used to study the effects of genes hypothesized to be important in the normal and neoplastic development of the human prostate (11). Recognizing the great utility of EST sequences for characterizing organ-specific gene expression, cloning novel genes and developing microarray reagent sets, we have initiated efforts to define the mouse prostate transcriptome by constructing and sequencing mouse prostate cDNA libraries. Interestingly, the extensive list of cDNA libraries provided at the Cancer Genome Anatomy Project web site lists more than 400 murine cDNA libraries, but none are derived from the prostate gland (http://www.ncbi.nlm.nih.gov/ncicgap/).

To date we have made five mouse prostate cDNA libraries, which are derived from microdissected anterior, dorsolateral and ventral prostatic lobes of mature mice, and from the urogenital sinus of E16 embryos. A total of 6145 ESTs have been sequenced, assembled, annotated and loaded into mPEDB in a fashion analogous to that described for processing human prostate sequence in PEDB. Virtual comparisons of transcriptomes derived from these distinct anatomical regions of the prostate suggest that the prostate lobes have specific functional attributes. Library summaries, text- and sequence-based queries, and virtual expression analyses tools are provided.

SUMMARY AND FUTURE DEVELOPMENTS

The human and mouse Prostate Expression Databases serve as centralized archives of gene expression information derived from the human and murine prostate that can be utilized by investigators studying normal and neoplastic prostate development. The assembled human prostate transcriptome currently comprises 20 187 distinct transcripts. Ongoing work involves the characterization of additional cDNA libraries representing specific prostate cell types and early developmental stages, the virtual comparative analyses of human and mouse prostate gene expression, and a database extension for archiving and analyzing cDNA microarray data derived from PEDB and mPEDB sequence resources. PEDB is accessible via the World Wide Web at http://www.pedb.org. mPEDB is accessible at http://www.mpedb.org or through a link from PEDB.

ACKNOWLEDGEMENTS

We thank Alec Mooreman, Stacy Boness and Josh Epstein for sequencing assistance. We thank Burak Eroglu for database administration. This work was supported by the CaPCURE Foundation and grants to P.S.N. from the NCI (K08 CA75173-01A1), NIDDK (RO1 DK59125-02) and the USAMRMC Program in Prostate Cancer Research (DAMD17-98-1-8499).

REFERENCES

1. Greenlee,R.T., Murray,T., Bolden,S. and Wingo,P.A. (2000) Cancer statistics. *CA Cancer J. Clin.*, **50**, 7–33.
2. Hawkins,V., Doll,D., Bumgarner,R., Smith,T., Abajian,C., Hood,L. and Nelson,P.S. (1999) PEDB: the Prostate Expression Database. *Nucleic Acids Res.*, **27**, 204–208.
3. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F.F. (1998) GenBank. *Nucleic Acids Res.*, **26**, 1–7. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 17–20.
4. Boguski,M.S., Lowe,T.M.J. and Tolstoshev,C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nature Genet.*, **4**, 332–333.
5. Schaefer,C., Grouse,L., Buetow,K. and Strausberg,R.L. (2001) A new cancer genome anatomy project web resource for the community. *Cancer J.*, **7**, 52–60.
6. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
7. Gordon,D., Abajian,C., Green,P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.*, **8**, 195–202.
8. Audic,S. and Claverie,J.M. (1997) The significance of digital gene expression profiles. *Genome Res.*, **7**, 986–995.
9. Greenberg,N.M., DeMayo,F., Finegold,M.J., Medina,D., Tilley,W.D., Aspinall,J.O., Cunha,G.R., Donjacour,A.A., Matusik,R.J. and Rosen,J.M. (1995) Prostate cancer in a transgenic mouse. *Proc. Natl Acad. Sci. USA*, **92**, 3439–3443.
10. Di Cristofano,A., De Acetis,M., Koff,A., Cordon-Cardo,C. and Pandolfi,P.P. (2001) Pten and p27KIP1 cooperate in prostate cancer tumor suppression in the mouse. *Nature Genet.*, **27**, 222–224.
11. Bhatia-Gaur,R., Donjacour,A.A., Sciavolino,P.J., Kim,M., Desai,N., Young,P., Norton,C.R., Gridley,T., Cardiff,R.D., Cunha,G.R., Abate-Shen,C. and Shen,M.M. (1999) Roles for Nkx3.1 in prostate development and cancer. *Genes Dev.*, **13**, 966–977.